INFERENTIAL REASONING: LEARNING TO "MAKE A CALL" IN PRACTICE

Maxine Pfannkuch
Department of Statistics, The University of Auckland, New Zealand
m.pfannkuch@auckland.ac.nz

Before students are introduced to formal statistical inference procedures at Grade 12, it is necessary in the earlier years to start building concepts such as sample, population, sampling variability, and "making a call". This paper reports on an initial study to determine whether proposed ideas about improving students' inferential reasoning were possible in practice in terms of student capability. Interviews, assessment responses, and classroom experiences will be used to describe how three low-to-average ability Grade 11 students started to conceptualize that inferences about populations can be made from samples. The findings suggest that the proposed instruction methods, which focus on building concepts about sampling variability, can influence and develop students' inferential reasoning. Issues arising from the study are discussed.

INTRODUCTION

Traditionally the teaching of formal statistical inference starts in Grade 12 and involves considering sampling distributions. But research continues to show that students' inferential understanding remains limited as they have a great deal of difficulty juggling many statistical objects: individual data values, samples, sample means, and distribution of sample means (Shaughnessy, 2007). Wild et al., (2009) believe that traditional ways of conceptualizing the sampling distribution requires learners to take their eyes off the sample plots and therefore the connections among sample, sampling distribution, statistics, and population become lost. Cobb (2007, p. 7) argues that teaching the sampling distribution is an "intellectual albatross" that is "remote from the logic of inference." He contends that the logic of inference must be placed at the heart of the statistics curriculum rather than the normal distribution. Placing inference at the centre of the secondary school curriculum necessitates a rethink on how to build students' inferential reasoning. There are many concepts associated with inference such as sample, population, distribution, and sampling variability. And how decisions are made about populations or processes from samples depends on "understanding, explaining and quantifying the variability in data" (Franklin et al., 2007, p. 6). Therefore the complexity of the concepts surrounding statistical inference suggests learning pathways need to be mapped out over several years of schooling.

In respect to mapping a learning pathway for inference that allows students to move "beyond the data in hand to draw conclusions about some wider universe, taking into account that variation is everywhere and the conclusions are uncertain" (Moore, 2007, p. xxviii), Wild et al. (2009) proposed a teaching sequence for comparative reasoning from Grade 9 to Grade 12. The main principles behind the instruction are:

- Learners from Grade 9 onwards should experience sampling behavior using direct dynamic *visual imagery* that strongly connects sampling variability, sample plots, and inference.
- There should be substantial hands-on simulation activity before moving to a computer environment (Shaughnessy, 2007).
- Clearly delineate descriptive and inferential thoughts (Pfannkuch et al., 2009).
- Enable students at these levels to "make a call" or decision about whether condition A tends to have bigger values than condition B.

This paper reports on an initial study to ascertain whether *some* of Wild et al.'s (2009) proposals would work in practice with low to average ability students in Grade 11.

THE PROBLEM

In New Zealand students sit national internal and external assessment standards in Grade 10, 11, and 12. In Grade 10 an internal statistics standard requires students to pose a comparison question from a given multivariate dataset, construct plots, analyze the data, and draw a conclusion, which must be justified with three pieces of evidence. Pfannkuch (2006) noted that for this standard the teacher and students did not know whether they were reasoning about the samples or about the populations from the samples. Pratt et al. (2008) report this confusion is common.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

Furthermore, Pratt and Ainley (2008) suggest that learners using informal inferential methods should be reasoning about populations from samples.

In Grade 11, an internal assessment statistics standard on *Sampling* requires students to select a sample and use this to make an inference about the population. The focus of the standard is on learning different sampling methods such as random, systematic, and stratified sampling, and once the sample is selected to calculate many sample statistics. Using the mean and standard deviation the students make an inference about the population. What is concerning is that the students do not plot their data and do not account for sampling variability or the sample size effect when they make an inference about the population mean. In fact in class they do not attend to or experience sampling behavior. Some publications in New Zealand even suggest that an interval for an inference about the population mean can be the sample mean plus or minus the sample standard deviation! This problem is not unique to New Zealand. Indeed, Saldanha and Thompson (2002, p. 268) stated: "in statistics instruction … it is uncommon to help students conceive of samples and sampling in ways that support their developing coherent understandings of *why* statisticians have confidence in this practice." Consequently, Wild et al. (2009) believe new approaches to learning statistical inference must be devised to assist teachers to improve students' learning of statistics.

RESOLVING THE PROBLEM AND THE RESEARCH QUESTION

Wild et al. (2009) and Pfannkuch et al. (2009) set out desired ways of reasoning comparatively, how to "make a call" at the different grade levels, and produced "movies" to show the effects of sampling variability. The problem was to produce hands-on simulations that moved seamlessly to the computer simulation "movies" and to define key learning experiences that students needed to develop their inferential reasoning and to "make a call." Since practitioners know what is feasible in a classroom environment, a research project team consisting of two statisticians, two statistics education researchers, and eight teachers was set up. In a series of meetings possible hands-on activities were proposed and then the team engaged in very intense and robust debates about introducing and articulating inferential ideas to students. From these discussions some key activities were designed to trial in some classrooms.

This paper reports on one of these trials. It focuses on the research question: What ideas do three Grade 11 students have about sample, population, sampling variability, and "making a call" before and after a teaching intervention?

METHOD

The research method is closely allied to design research (Roth, 2005). In this case, however, the proposed teaching sequence for the new curriculum had to be modified to incorporate the demands of the current assessment system for Grade 11. The teacher involved was a member of the research project team. Since she was responsible for getting the students ready to sit the internal assessment *Sampling* standard, she decided what activities she would use from the project and how far she could extend the students beyond what was currently required within the time available. Her aim was to introduce students to the following ideas: sample, population, sampling variability, sample size effect, making an inference about a population from a sample using an informal confidence interval for the population median (Figure 1), and expressing the level of confidence in capturing the true population median. The researcher was present in her classroom and there was interaction between them about the lessons.

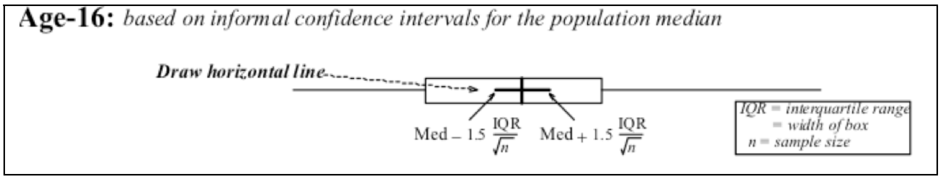


Figure 1. Initial basis for how to make a call at Grade 11 or age 16 (Wild et al., 2009)

In order to understand whether the proposed teaching approaches have the potential to improve students' inferential reasoning, a case study involving three students will be the focus of this analysis. The rationale for the approach is that an in depth analysis will give some invaluable insights and information into students' reasoning processes. The students were in a high socio-economic multicultural girls' school. Because they were classed as low-to-average-ability in mathematics they were not doing a course that would advance them to Grade 12. In the previous and current year they had sat internal assessment standards involving the comparison of box plots.

Before the unit was taught all the students in the class sat the pre-test, consisting of three tasks, and then the researcher individually interviewed and probed the reasoning of the three case-study students about their responses to the pre-test. The class lessons that related to inferential reasoning were videotaped. About twelve days after the lessons relating to inference were taught the students in the class sat the post-test, in which one task was the same as the pre-test. The case study students were interviewed again. A qualitative analysis of the interviews and tests of the three case-study students was conducted. Before reporting on these students' reasoning, a brief description of two key learning experiences will be given as well as some brief reflections on implementation.

Some classroom experiences and reflections

The first key learning experience was the KiwiKapers activity, which involved a population of 730 kiwis in a bag. Each kiwi, a New Zealand native bird, was a square datacard, which contained the following information: species, weight, height, location, and gender. The class discussed questions they could pose about the kiwi population. The teacher then suggested they would all pose the question: "What is the typical weight of a kiwi?" Groups of two students had a bag, from which they randomly drew a sample of 30 kiwis or datacards. On the paper scale they were given they constructed datacard plots and box plots (Figure 2). Since the aim of the lesson was to explore sampling variability and its *properties*, the whiskers were left off. In earlier research Pfannkuch (2008) noted the presence of outliers sidetracked students' reasoning and the research project team independently decided the focus should be on the boxes. Colors for the median and boxes were deliberately kept the same as the "movies." Once all the plots were pinned on the wall, students were asked what they noticed. The extent of the variability in the medians for samples of size 30 was plotted. In order to understand the effect of sample size, students repeated this activity for samples of size 100, but this time summary statistics were derived from computer simulations. Computer simulation "movies" then further reinforced their visual imagery of sampling variability.

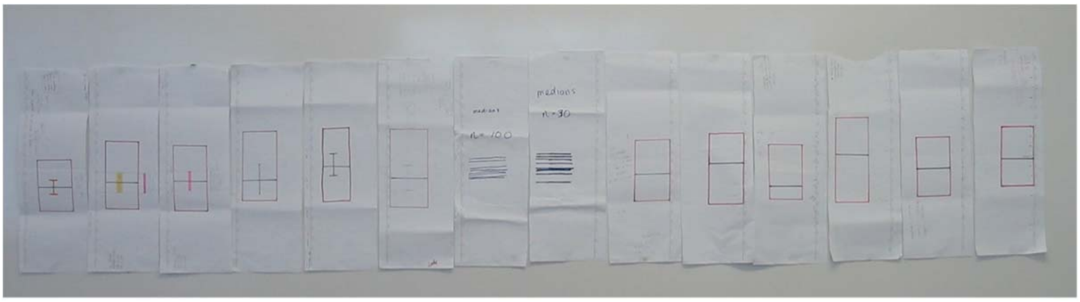


Figure 2. Wall display showing sampling variability and sample size effect for n=30 and n=100

The second key learning experience was to calculate the confidence interval (C.I.) for the population median and determine how often it captured the true population median. At this point there are two problems. The first problem is whether the students will understand the formula for the C.I. They may understand why *n* is part of the formula but not the interquartile range (IQR) since they have not experienced a key learning experience in Grade 10, which shows the IQR needs to be taken into account when making an inference. The second problem occurred when the teacher got the students to draw their C.Is for their plots of n=30 and n=100 on some new paper scales to see how often their C.Is captured the true population median. By using new paper scales rather than the ones in Figure 2 she violated a Wild et al. (2009) principle that sample plots and sampling variability properties must stay visually connected. It should also be noted there was confusion amongst research project team members about sample sizes and confidence levels for capturing the true population median.

RESULTS

Students' reasoning before the teaching intervention

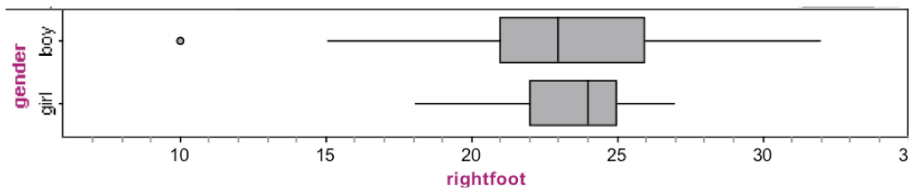
The three students were interviewed about their responses to the pre-test and were asked further questions to probe their reasoning. The task in Figure 3 was designed to probe their reasoning and ideas about sample and population, sampling variability, and "making a call."

Emma takes a *random sample* of 30 Year 8 NZ boys and a *random sample* of 30 Year 8 NZ girls. She can find out their right foot lengths without shoes on in cm.

1. Emma knows that her *random sample* of 30 Year 8 NZ boys can be used to find out …

Emma is interested in comparing the right foot lengths (in cm) of Year 8 NZ boys and girls. She plots her graph correctly.

Emma's graph



Emma looks at her graph and claims that the right foot lengths of Year 8 NZ girls tend to be bigger than the right foot lengths of Year 8 NZ boys.

2. Would you make the same claim as Emma? Why?

Figure 3. A pre and post-test task: Questions 1 and 2

These students had no idea that random samples were taken to make inferences about the population. For Question 1 all stated that the random sample could be used to find out the average right foot length. S1 and S2 were very clear that the average was related to the data in hand:

S1: You get a whole set of data … you can find out the highest and the lowest, and the upper and lower quartiles, and the average … in statistics the average is normally used to generalize the heights.
I: So you are saying that she can find the average foot length *just for those* boys.
S1: Yes
I: So she can't find out that out *for all* Year 8 boys in NZ?
S1: I don't know if by using those 30 boys you could say it's for all Year 8 boys. I don't know if you could do that.
I: Why did she take a random sample?
S2: Just to see a different kind of variety of different foot lengths and then to find the average I guess.

S3, on the other hand, argued that a sample of size 30 could tell you about the average foot length of all Year 8 NZ boys, because that is all she knew and what she had been taught. In fact she would take a sample of size 30 in every given situation. Despite a lot of probing there was no evidence she understood the relationship between sample and population.

For Question 2, S1 agreed with Emma and "made a call" based on the medians but when probed she said "Like if this tail to the right [for the girls] went out further then may be I would have agreed more but I don't know whether you can just go by the median." S2 and S3 disagreed with Emma and "made a call" on the fact that the IQR for the boys' foot length was larger and the UQ for the boys' foot length was higher than the girls'. There was no rationale behind their choices of measure for "making a call" but it seemed to be based on an idiosyncratic view of the "picture".

To ascertain their concept of sampling variability they were asked what the graph would look like if another person took a random sample of 30 Year 8 boys and 30 Year 8 girls, and then to draw the graph. Intuitively they knew that if another sample were taken the results would be different. But their thinking was very akin to the outcome approach (Konold, 1989). As S3 stated: "I mean honestly, it could just be anything." S3 believed that since Year 8 students were growing, their foot lengths in another sample could be quite different but if they were adults the two graphs would be similar. When they drew their graph S2 kept the medians in the same direction, S1 put the medians the same as she thought boys' foot lengths should be bigger and S3 only drew a graph for adults with the male foot lengths shifted much further along than the females'. Basically they did not know the extent of sampling variability or that the direction of the shift including the sample median could be reversed. Context knowledge and the outcome approach to chance seemed to play a part in their thinking about sampling variability.

Students' reasoning after the teaching intervention

After the teaching intervention, in the interviews, the three students were able to articulate sample and population ideas.

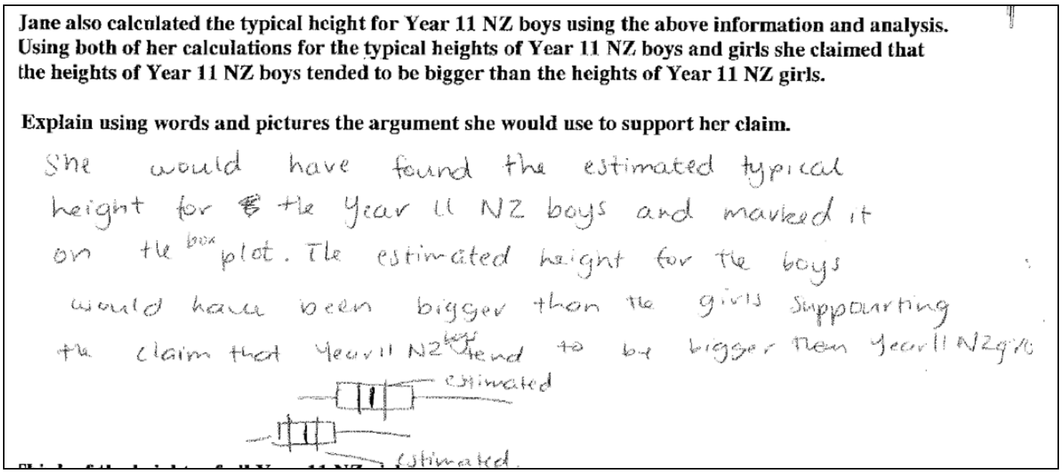
S1: You're just getting a sample of the whole population and then you can, by that sample, you can look at it and make a statement or whatever about the whole population. You can make a pretty accurate estimate of the whole population median just from taking a sample of, even, thirty.

The students were able to *articulate* and *draw images* for the extent of sampling variability in the medians for samples of size 30 and 100. Their responses indicated they could remember the Figure 2 image on the wall but found it difficult to recall the C.I. images, also on the wall. All had an image of a box with variability in the sample median similar to the "movie" images.

In the post-test they were asked to find the typical height of Year 11 NZ girls for a sample of size 32, then they were asked how another student's answer would differ from their one if she took a sample of 100. S1 wrote: "The range of possible population medians would be smaller, meaning the estimate will be more accurate." A problem area, though, was the confidence level. S1 and S2 knew the confidence level for a sample of size 30 was 90% but thought it was much higher for 100. S3, however, when responding to further questions about Question 1 (Figure 3) said:

S3: It's a sample of size 30, I'd say nine times out of 10.
I: What makes you believe that?
S3: When we did comparing with the kiwis, we had the actual median for the box plot, and then we did our sample and then we found, like the range of medians that it could be and whether it crossed over the actual median line or not and nine times out of 10 it did. So that's why.
I: What if she took a bigger sample?
S3: If she took a bigger sample then the median range would be smaller, so more accurate results, like more confined. …
I: And how confident would you be [about capturing the true population median] if she took a 100.
S3: I'd still be nine times out of ten.

Since the assessment *Sampling* standard did not require students to compare two groups and "make a call", this was not part of the teaching unit. In the post-test students were given a graph comparing the heights of Year 11 NZ boys and girls, for which they were asked to find the typical height of girls. Another question probed whether they could "make a call" based on their current knowledge. As S2 shows in Figure 4 she was very close to being able to "make a call."


Jane also calculated the typical height for Year 11 NZ boys using the above information and analysis. Using both of her calculations for the typical heights of Year 11 NZ boys and girls she claimed that the heights of Year 11 NZ boys tended to be bigger than the heights of Year 11 NZ girls.

Explain using words and pictures the argument she would use to support her claim.

She would have found the estimated typical height for ~~B~~ the Year 11 NZ boys and marked it on the box plot. The estimated height for the boys would have been bigger than the girls supporting the claim that Year 11 NZ boys tend to be bigger than Year 11 NZ girls

Figure 4. Post-test response of student S2

In summary, prior to the teaching intervention these three students did not seem to have any understanding that they could reason about a population from a sample or have any visual imagery associated with sampling variability. Some rudimentary ideas about random samples being different were present. After the teaching intervention these students appeared to have some concepts, including visual imagery, about sample, population, and sampling variability for quantitative data displayed as a box plot. They also seemed to be on the verge of understanding how to "make a call" for comparing two box plots.

CONCLUSION

The purpose of this paper was to ascertain whether some of Wild et al.'s (2009) proposals for embedding statistical inferential ideas into students' reasoning would work in practice. From the responses of these three low-to-average ability Grade 11 students, it would seem that some of their proposals are viable. These students were beginning to conceive that samples could provide information about a population and that sampling variability needed to be taken into account when drawing conclusions. Although "making a call" or making a decision in a comparative situation was not part of the teaching unit, I conjecture the conceptions they were beginning to articulate and visualize would help them make that transition. This initial study focused on some concepts related to inference (Cobb, 2007) and variability (Franklin et al., 2007). However, as Shaughnessy (2007, p. 981) stated: "The concepts surrounding statistical inference are very complex, and the transition for students is likely to be in process for several years. There is no quick fix for understanding these concepts." What is important is that students should begin to be immersed in inferential ideas from at least Grade 9 before being introduced to sampling distributions at Grade 12. Researchers also need to build a knowledge base on the teaching and learning of statistical inference.

ACKNOWLEDGEMENT

This work is supported by a grant from the Teaching and Learning Research Initiative. Thanks to Chris Wild for his helpful comments on this paper.

REFERENCES

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1). Online: http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: a pre-k-12 curriculum framework*. Alexandria, VA: American Statistical Association.

Konold, C. (1989). Informal Conceptions of Probability. *Cognition and Instruction*, *6*(1), 59-98.

Moore, D. (2007). *The basic practice of statistics* (4th edition). New York: W. H. Freeman and Company.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27-45. Online: www.stat.auckland.ac.nz/serj.

Pfannkuch, M. (2008). Building sampling concepts for statistical inference: A case study. In *ICME-11 Proceedings, Monterrey, Mexico, July 2008*. Online: http://tsg.icme11.org/tsg/show/15.

Pfannkuch, M., Regan, M., Wild, C. J., & Horton, N. (2009). *Telling data stories: essential dialogues for comparative reasoning*. Unpublished paper.

Pratt, D., & Ainley, J. (2008). Introducing the special issue on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 3-4. Online: www.stat.auckland.ac.nz/serj.

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, *7*(2), 107-129. Online: www.stat.auckland.ac.nz/serj

Roth, W. (2005). *Doing qualitative research: praxis of method*. Rotterdam: Sense Publishers.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*, 257-270.

Shaughnessy, M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second handbook of research on the teaching and learning of mathematics*, (Vol. 2, pp. 957-1009). Charlotte, NC: Information Age Publishers.

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2009). *Precursor statistical inferences*. Unpublished paper.